# Homework # 2

The data in the *Streams* sheet of the *data\_mrm.xlsx* Excel workbook present measurements of electroconductivity of stream water at multiple sites along two water streams (*L* and *U*). Beyond stream identity (variable *Stream*), the variable *Velocity* represents water speed at the collection spot.

Your main task is to develop a linear model (with *EC* and *Velocity* used with a log-transformed scale), which predicts *EC* values using *Velocity* and *Stream* variables as predictors. Naturally, if any of the variables is not worth of keeping in the model, you should drop it. Additionally, if at least one of the predictors is significantly related to *EC*, you should also try to add the interaction of both predictors.

Do not just fit the appropriate models and perform appropriate statistical tests: describe also what you have done (used methods) and what the results were - using whole sentences, as in a research paper.

Further, you should examine the effects of predictor(s) in the resulting model and describe them using whole sentence including a quantitative description what the predictor' effect size is (e.g. "If the stream velocity doubles, the EC value changes by ..." - but only if such effect remained in the model, for sure).

Place the used commands (and their results) together with the answers to the above questions below in a Word document with the name *Surname*.docx and submit it to Moodle page before Friday 23:59.

I look forward to your solution.

Petr Šmilauer

**METHODS**

First of all, EC and Velocity values are log transformed. It means that a difference of 1 on the log scale corresponds to a doubling of EC. In the following text, when EC and Velocity are mentioned, it in fact refers to their log transformed values.

A close up of a text

Description automatically generated

1. **Pre-Analysis : Visual representation of the correlations between all variables**

A splom plot is created so as to check what predictors seem to have the strongest relationship with EC and how the predictors are correlated with each others.

1. **Determination of the best linear model to predict EC**

The more parsimonious model explaining EC variation is found and the summary of the best model is used to know in what direction the predictors affect EC.

1. **Description of the effects of the predictors and graphical representation**

A plot of the relationship between the predictors and EC is realized so as to illustrate the model summary.

**RESULTS**

1. **Pre-Analysis : Visual representation of the correlations between all variables**



**A grid of multiple scatter matrix

Description automatically generated***Figure 1: Splom plot of all the variables of the dataset*

EC and Velocity seem to have no relationship (datapoints distributed randomly) while there seem to be a linear relationship between EC and Stream. There is also no trend between Velocity and Stream, they seem to be independent from each other.

1. **Determination of the best linear model to predict EC**

First, we build a full model with all predictors: EC ~ Velocity + Stream.



A screenshot of a computer

Description automatically generated

*Figure 2: Full model with all predictors*

Stream is the only predictor which has a significant effect on EC. The model explains a lot of variation with R2=0.9254, and the power of the model is strong: lm(F2,37 =242.9 ; p=2.2e-16).

Then, I build the model with the interaction Velocity \* Stream to check its effect on EC.



A screenshot of a computer program

Description automatically generated

*Figure 3: Full model with all predictors in interaction*

The interaction Velocity \* Stream has no significant effect on EC, only Stream is affecting significantly EC. The variance explained by this model is even lower than the previous one: R2 = 0.9248. The power of the model is lower with lm(F3,36=160.9 ; p=2.2e-16).

The best model seems then to be a model with only Stream as a predictor of EC.

A close up of a logo

Description automatically generated

A screenshot of a computer program

Description automatically generated

*Figure 4: Best model to predict EC*

This model is indeed the best, it has the highest explained variation with R2 = 0.9258 and the best power with the highest F-value : lm(F1,38=487.3 ; p=2.2e-16).

Finally, the *add* function is used to double check if adding Velocity in the model would make it better, which should not be the case based on what we just saw previsouly.



A screenshot of a computer code

Description automatically generated

*Figure 5: Check with the add function what changes in the model if we add the predictor Velocity*

We can see that adding Velocity as a predictor in the model does not make the AIC lower and does not lower the p-value (it does not add any significancy). We made the right choice by chosing the model with the single predictor Stream.

It can be noted that I also checked the significancy of the effect of the predictors with the Anova type III (non sequential Anova). It gave the same results.

Finally, the validity of the best model (*Figure 4*) has been verified.

A white background with black text

Description automatically generated

A group of graphs with numbers

Description automatically generated with medium confidence

*Figure 6: Check of the vailidity of the model*

The model meets the assumption of normality of the residuals and of homogeneity of variances (homoscedasticity).

1. **Description of the effects of the predictors and graphical representation**

**A close up of words

Description automatically generated**

A diagram of a graph

Description automatically generated

*Figure 7: Graphic of the effect of Stream on EC*

Based on the model summary (*Figure 4*) and the graphic above, if there is a change in one unit in Stream (which means going from Stream L to Stream U), the EC value increases by 1.16322 (=estimate of Stream effect in the final model), which means that EC increases from 16% between Stream L and Stream U.

**CONCLUSION**

Water speed does not affect the electroconductivity of the water, only the site does: the water has a significant higher electroconductivity at the Stream U than at the stream L.